Logistic Regression

Assumptions, Limitations, and Applications

Ahmed Badi

Mathematics & Machine Learning Enthusiast

ahmedbadi905@gmail.com

December 5, 2025

Outline

Introduction to Logistic Regression

The Classification Problem

Definition

Logistic regression predicts the probability that an observation belongs to a binary class

Examples

- Spam / Not spam
- Disease / Healthy
- Click / No click
- Fraud / Legitimate

Key Features

- Probabilistic output [0, 1]
- Interpretable coefficients
- Efficient and scalable
- Foundation for many techniques

The Logistic Function

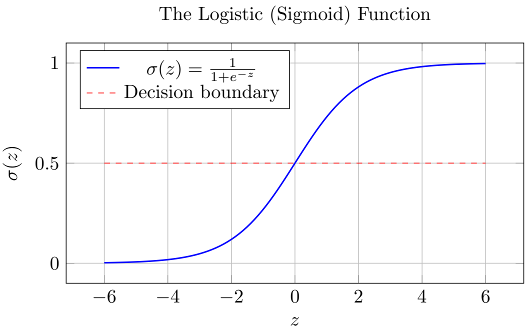
The Sigmoid Function

Logistic Function

$$\sigma(z) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}$$

Key Properties:

- Range: always in (0, 1)
- Smooth and differentiable
- S-shaped (sigmoidal) curve
- Symmetric around 0.5



Logistic Regression Model

Simple Model (One Predictor)

$$p(y = 1 \mid x) = \sigma(\beta_0 + \beta_1 x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

Multiple Predictors

$$p(y = 1 \mid \mathrm{x}) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}}$$

Vector Form

$$p(y = 1 \mid \mathrm{x}) = \sigma(\beta^\top \mathrm{x})$$

where $\beta = [\beta_0, \beta_1, \ldots, \beta_k]^\top$ and $\mathrm{x} = [1, x_1, \ldots, x_k]^\top$

Log-Odds and Interpretation

Odds

$$\text{Odds} = \frac{p(y=1)}{p(y=0)} = \frac{p(y=1)}{1-p(y=1)}$$

Log-Odds (Logit)

$$\log\left(\frac{p(y=1)}{1-p(y=1)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Interpretation

A unit increase in x_1 multiplies the odds by e^{β_1} (odds ratio)

Maximum Likelihood Estimation

Likelihood and Log-Likelihood

Bernoulli Likelihood

For observation i:

$$L_i(\beta) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

where $p_i = \sigma(\beta^\top \mathrm{x}_i)$

Full Log-Likelihood

$$\ell(\beta) = \sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Goal

Maximize $\ell(\beta)$ to find optimal parameters $\hat{\beta}$

Cost Function: Cross-Entropy Loss

Binary Cross-Entropy

$$J(\beta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \log(\hat{p}_i) + (1 - y_i)\log(1 - \hat{p}_i)\right]$$

Gradient for Optimization

$$\frac{\partial J}{\partial \beta_j} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)x_{i,j}$$

No Closed Form

Unlike linear regression, no analytical solution exists. Use gradient descent or Newton's method.

Optimization Methods

Gradient Descent

- Iterative approach
- Simple, stable
- Slower convergence

Newton's Method

- Faster convergence
- Quadratic convergence
- More computationally intensive

Quasi-Newton

- Balance of both
- (BFGS, L-BFGS)
- Practical choice

Assumptions and Limitations

Key Assumptions

Model Assumptions:

1. Binary outcome (0/1)
2. Linear log-odds
3. Independence of observations
4. No multicollinearity

Data Assumptions:

1. Sufficient sample size
2. No extreme outliers
3. Correct model specification
4. No omitted variables

Violations and Remedies

Multicollinearity

- Check VIF (variance inflation factor)
- Remedy: regularization or feature selection

Linearity of Log-Odds

- Add polynomial/interaction terms
- Use kernel methods or GAMs

Class Imbalance

- Adjust class weights
- SMOTE, undersampling
- Threshold adjustment

Outliers

- Robust regression approaches
- L1 regularization

Model Evaluation

Confusion Matrix and Metrics

	Predicted 1	Predicted 0
Actual 1	TP	FN
Actual 0	FP	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

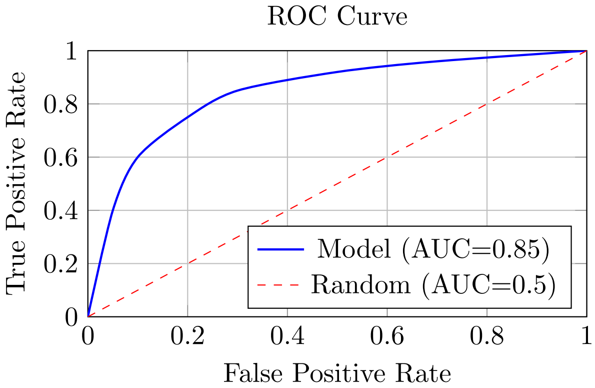
$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC Curve and AUC

ROC: Receiver Operating Characteristic

Plot True Positive Rate (sensitivity) vs False Positive Rate (1 - specificity) across all thresholds



Threshold Selection

Default Threshold

Standard: predict 1 if $\hat{p} > 0.5$, else predict 0

Problem

May not be optimal for imbalanced data or different cost of errors

Solution

Adjust threshold based on business needs:

- Low threshold $\rightarrow$ higher recall, lower precision
- High threshold $\rightarrow$ lower recall, higher precision

Regularization

L1 and L2 Regularization

L2 Regularization (Ridge)

$$J(\beta) = -\frac{1}{n}\sum y_i \log(\hat{p}_i) + \frac{\lambda}{2n}\sum \beta_j^2$$

Shrinks coefficients toward zero

L1 Regularization (Lasso)

$$J(\beta) = -\frac{1}{n}\sum y_i \log(\hat{p}_i) + \frac{\lambda}{n}\sum |\beta_j|$$

Can drive coefficients exactly to zero (feature selection)

Elastic Net

Combines L1 and L2 penalties for balance between regularization and feature selection

Extensions

Multinomial Logistic Regression

Multi-class Problem

Extends binary logistic regression to $K > 2$ classes

Softmax Function

$$p(y = k \mid \mathrm{x}) = \frac{e^{\beta_k^\top \mathrm{x}}}{\sum_{j=1}^{K} e^{\beta_j^\top \mathrm{x}}}$$

Key Features

- Probabilities sum to 1 across classes
- Same optimization (cross-entropy loss)
- Natural extension of binary case

Ordinal Logistic Regression

Use Case

When classes have natural ordering: poor $\rightarrow$ fair $\rightarrow$ good $\rightarrow$ excellent

Proportional Odds Model

Assumes log-odds change linearly across thresholds:

$$\log\left(\frac{p(y \leq j)}{p(y > j)}\right) = \alpha_j - \beta^\top \mathsf{x}$$

Advantage

Preserves ordering information; fewer parameters than multinomial

Advantages and Limitations

Advantages

Theoretical

- Well-understood, mature method
- Founded on solid statistics
- Probabilistic output
- Simple, elegant solution

Practical

- Fast to train and predict
- Works with small-medium data
- Interpretable coefficients
- Baseline for comparison

Limitations

Model Assumptions

- Assumes linear log-odds
- Requires correct specification
- Sensitive to multicollinearity
- Can underfit complex patterns

Data Issues

- Struggles with imbalance
- Affected by outliers
- Needs sufficient sample size
- Limited to structured data

Conclusion

Key Takeaways

1. Logistic regression transforms classification into probability estimation
2. Sigmoid function ensures valid probabilities [0, 1]
3. Maximum likelihood provides principled estimation
4. Interpretable coefficients via odds ratios
5. Regularization prevents overfitting
6. Rich evaluation metrics for assessment
7. Extends naturally to multinomial and ordinal settings

When to Use Logistic Regression

Good Fit

- Binary/multiclass outcome
- Need interpretability
- Moderate-sized datasets
- Baseline benchmark
- Regulatory requirements

Consider Alternatives

- Highly non-linear data
- Severe class imbalance
- Unstructured data
- Complex interactions
- Deep learning problems

Thank You!

Questions & Discussion

Ahmed Badi

Logistic Regression: Assumptions, Limitations, Applications

ahmedbadi905@gmail.com

December 5, 2025

Backup Slides

Handling Class Imbalance

Problem

Most data are class 0, few are class 1 $\rightarrow$ model biased toward majority class

Solutions:

1. Adjust class weights in cost function
2. SMOTE (Synthetic Minority Oversampling)
3. Undersampling majority class
4. Threshold adjustment
5. Stratified cross-validation

Feature Scaling for Logistic Regression

Standardization

$$x' = \frac{x - \mu}{\sigma}$$

Mean 0, standard deviation 1

Why Helpful

- Faster convergence of optimization
- Comparable coefficient magnitudes
- Better numerical stability
- Important if using L1/L2 regularization

Resources and References

- Hosmer, D. W., & Lemeshow, S. (2013). Applied Logistic Regression
- Cox, D. R. (1958). The Regression Analysis of Binary Sequences
- James, G., et al. (2021). Introduction to Statistical Learning
- Agresti, A. (2018). Categorical Data Analysis
- Andrew Ng, Machine Learning Course (Coursera)

Contact

Author: Ahmed Badi

Email: `ahmedbadi905@gmail.com`

LinkedIn: linkedin.com/in/badi-ahmed

Date: December 5, 2025

Open to collaborations